基于用户评分时间改进的协同过滤推荐算法*

李道国[1] 李连杰[2] 申恩平[2]
[1](杭州电子科技大学信息工程学院 杭州 310018)
[2](杭州电子科技大学管理学院 杭州 310018)

摘要:【目的】改进基于用户的协同过滤算法以缓解因数据稀疏、用户共同评分稀少所导致的问题, 进而提高评分预测的精度。【方法】提出结合用户打分时间发现具有相似打分行为的用户, 并将用户评分方差相似性融入到相似度的计算中, 使得目标用户在最近邻的选取上更加合理。【结果】实验结果表明, 相较基于用户的协同过滤算法, 新算法的平均绝对误差降低约 2%, 在一定程度上改善了推荐系统的推荐效果。【局限】该算法仅在 MovieLens 数据集上进行了实验测试, 还需要在其他数据集上进行检验。【结论】本文算法能够有效地提高推荐精度, 具有一定的可行性和现实意义。

关键词: 协同过滤 数据稀疏 相似评分 用户评分方差相似性 最近邻
分类号: TP311 G25

1 引 言

互联网的迅速发展将人们带入了一个崭新的信息时代, 网络中的信息资源越来越丰富, 当用户面对海量的数据信息时, 如何在茫茫的信息海洋中快速、准确地找到需要的信息成为用户关注的问题, 潜在的用户也常常因此而流失, 这就是所谓的“信息过载”现象[1-2]。为了使用户能够在庞大的数据中快速找到需要的信息, 个性化推荐应运而生。协同过滤推荐算法是其中应用最广泛的技术, 其优点是对所推荐的项目没有特别的要求, 而且还能够处理非结构化复杂的对象, 如文章、电影以及书籍等。协同过滤推荐算法通过分析用户–项目评分矩阵, 在此基础上将大量不需要的信息过滤掉, 最后寻找到用户所感兴趣的项目[3]。

虽然协同过滤推荐算法在很多方面表现出独特的优势, 但主要的缺点是过分依赖评分矩阵。随着网站商品和用户数量快速地增长, 评分矩阵中用户真正给予评分的商品数量非常少, 通常在 1%以下。当数据过于稀疏时, 推荐系统中用户之间的共同评分项目就会极其稀少, 这种情况使得用户之间的相似度计算不准确, 从而导致推荐质量下降。因此本文提出一种基于用户评分时间改进的协同过滤推荐算法, 该算法能够有效地缓解数据异常稀疏、用户共同评分稀少所带来的问题, 通过优化最近邻查找的方法, 提高推荐准确性。

2 研究背景

已有很多学者针对如何改善数据稀疏性对推荐系统的影响进行了大量的研究, 基本可分为两类: 利用一定的方法降低数据的稀疏度;改进推荐系统的推荐算法来提高算法的推荐质量。对于推荐算法的改进研究, 由于寻找目标用户的最近邻是协同过滤算法的核

通讯作者: 李连杰, ORCID: 0000-0002-5342-5366, E-mail: 985214427@qq.com。

*本文系浙江省自然基金项目“技术知识特性、整合、知识能量与组织学习对企业间合作创新能力关联性研究”(项目编号: LY12G01002)的研究成果之一。

心，在推荐效果上起着至关重要的作用[4]，因此用户之间相似度计算的准确性就非常关键。传统的相似度计算方法主要有：余弦相似性、修正余弦相似性、Pearson 相关相似性[5-6]。采用余弦相似性计算用户之间相似度的过程中没有充分利用评分时间这一信息，评分时间对于判断该项目评分的有用性具有关键的作用。虽然修正的余弦相似度和 Pearson 相关相似性改善了评分分值所带来的影响，但也没有考虑到一种相对特殊的情况，即当评分矩阵异常稀疏时，两个用户之间的共同评分的项目就会极少，使用该方法计算的相似度同样会存在不准确的情况，导致推荐质量效果不佳的问题。

针对上述问题，本文提出结合用户的评分时间发现具有相似评分行为的用户，从而改善传统协同过滤算法中寻找最近邻的方法。并在此基础上融合用户评分方差相似度，从而更全面地利用用户评分信息改善相似度的计算，即使在数据异常稀疏、用户之间共同评分稀少的前提下依然能相对准确地计算用户之间的相似度，达到提高推荐准确性的目的。

3 基于用户评分时间改进的协同过滤推荐算法

3.1 算法描述

(1) 定义 1：相似评分项，T_{ui} 表示用户 u 对项目 i 的评分时间，假设用户 u 和 v 都对项目 i 有过评分，另外 T_{ui} 与 T_{vi} 的差值小于预先指定的一个时间间隔，那么 i 就是被认定为用户 u 和 v 的相似评分项目 S_{uv}。

(2) 定义 2：相似用户行为，两个用户的相似评分项大于等于指定阈值 λ，则认为这两个用户具有相似用户行为，公式如下：

$$S=\begin{cases}0 & S_{uv}<\lambda \\ 1 & S_{uv}\geqslant\lambda\end{cases} \tag{1}$$

当 $S_{uv}<\lambda$ 时，表示用户 u 和用户 v 之间相似评分行为少，在计算相似性时需要舍弃，否则会影响准确性。用户相似评分项阈值 λ 在实验中的取值直接影响到算法计算的准确性，在应用的过程中需要根据实验中的具体情况确定 λ 的取值，从而得到最优解。

(3) 定义 3：用户评分方差相似度

本文将用户的评分方差引入到相似度的计算中以衡量用户之间相似度的差异性，在此基础上提出用户评分方差相似度(User Rating Variance Similarity, URVS)理论，计算方法如公式(2)所示：

$$Sim_{URVS}(u,v)=1-\frac{|Var_u-Var_v|}{Var_u+Var_v} \tag{2}$$

其中，Var_u，Var_v 分别表示用户 u, v 的评分方差，例如用户 a, b, c 的方差分别为 1、3、5，那么 $Sim_{URVS}(a,b)=0.5$，$Sim_{URVS}(a,c)\approx 0.33$。评分方差越大表示用户的争议度就越大，评分方差越小表示用户的争议度就越小。

3.2 改进后的相似度计算

本文以修正的余弦相似度为例，结合用户评分时间的相似度的计算方法如公式(3)所示：

$$sim(\vec{u},\vec{v})_{A-c}=S\times\frac{\sum_{i\in I_{uv}}(R_{ui}-\bar{R}_u)(R_{vi}-\bar{R}_v)}{\sqrt{\sum_{i\in I_u}(R_{ui}-\bar{R}_u)^2\sum_{i\in I_v}(R_{vi}-\bar{R}_v)^2}} \tag{3}$$

在结合用户评分时间的基础上，引入用户评分方差相似度的相似性计算如公式(4)所示：

$$Sim_{URVS-CF}(u,v)=\alpha Sim_{URVS}(u,v)+(1-\alpha)Sim_{A-c}(u,v)\times S \tag{4}$$

即：

$$Sim_{URVS-CF}(u,v)=\alpha\left(1-\frac{|Var_u-Var_v|}{Var_u+Var_v}\right)+(1-\alpha)\frac{\sum_{i\in I_{uv}}(R_{ui}-\bar{R}_u)(R_{vi}-\bar{R}_v)}{\sqrt{\sum_{i\in I_u}(R_{ui}-\bar{R}_u)^2\sum_{i\in I_v}(R_{vi}-\bar{R}_v)^2}}\times S \tag{5}$$

基于用户评分时间改进的协同过滤推荐算法的优势在于即便在数据异常稀疏、用户之间共同评分稀少的情况下依然能够充分地利用用户的评分信息，相对准确地计算出用户之间的相似度，提高推荐系统的推荐准确性。

3.3 改进后算法的主要步骤

输入：用户−项目评分矩阵，目标用户 u

输出：目标用户 u 的 TOP-N 项目推荐列表。

①根据用户−项目评分矩阵 R，利用改进的修正余弦相似度计算方法(公式(4))计算用户 u 和其他用户的相似度，如果在某个时间段内评价电影的个数过少，则相似度 Sim 设为 0。

②根据步骤①计算出的相似度，确定目标用户 u 的 k 个最近邻居，设最近邻居集合为 $K=\{v_1,v_2,\cdots,v_k\}$，则目标用户 u 与最近邻的相似度为 $sim=\{sim_{u1},sim_{u2},\cdots,sim_{uk}\}$。

③分别确定目标用户 u 和相似近邻已经评分过的项目集合 I_u 和 $I_i=\{i_1,i_2,\cdots,i_k\}$，将所有的 I_i 取并集，然后将 I_u

中已经存在的项目去掉，最后产生候选集 Z。

④对候选集中 $\forall j \in Z$，利用公式(6)预测用户 u 对项目 j 的评分。

$$P_{ui} = \bar{R}_u + \frac{\sum_{v_i \in N} sim_{URVS-CF}(\vec{u},\vec{v}) \cdot (R_{vi} - \bar{R}_v)}{\sum_{u_i \in N} |sim_{URVS-CF}(\vec{u},\vec{v})|} \quad (6)$$

⑤将步骤④中项目的预测评分按照降序从大到小的排列，选择排在最前面评分最高的前 n 个项目推荐给用户 u。

4　实验与分析

4.1　实验数据与环境

本文采用由 Minnesota 大学的 GroupLens 研究小组创建的 MovieLens 数据集①中的100K 的数据集进行实验。该数据集记录共943个用户对1 682部电影的10万条评分。评分的分值在[0-5]之间不等，用户对电影的喜爱程度随着分值的增加而递增[7]。从数据集中随机抽取80%作为训练集，剩余20%作为测试集[8]。数据集的稀疏度计算如下所示：

$$\varphi = 1 - \frac{100000}{943 \times 1682} \approx 0.936$$

由此可见，所选择的数据集的评分矩阵是非常稀疏的。

实验环境是 Intel(R)Core(TM)i3-2310M 2.10GHz CPU，2GB 内存，Microsoft Windows7 操作系统，算法使用 Matlab 语言实现。

4.2　检验指标

采用平均绝对误差(MAE)评价该系统的推荐质量。MAE 通过计算实际评分与预测评分之间的差值衡量算法的好坏。MAE 值越小说明该算法就越好。MAE 值的计算如公式(7)所示：

$$MAE = \frac{1}{n} \sum_{(u,i) \in R} |P_{u,i} - R_{u,i}| \quad (7)$$

其中，$P_{u,i}$ 表示用户 u 对电影 i 的预测评分，$R_{u,i}$ 表示用户 u 对电影 i 的真实评分，n 表示 $P_{u,i}$ 或者 $R_{u,i}$ 的数量。

4.3　实验分析

(1) 实验一：参数 λ 对推荐系统性能的影响

利用公式(3)计算出用户之间的相似度，并根据相似度的大小确定目标用户的最近邻，根据文献[6]中的评分预测公式计算出目标用户对于未评分项目的评分值。由于实验一主要用于测试参数 λ 对 MAE 值的影响，因此需要控制最近邻的个数，当最近邻的个数为 30，阈值的取值分别是 3、5、8、10、12、14, MAE 值的变化情况如图 1 所示：

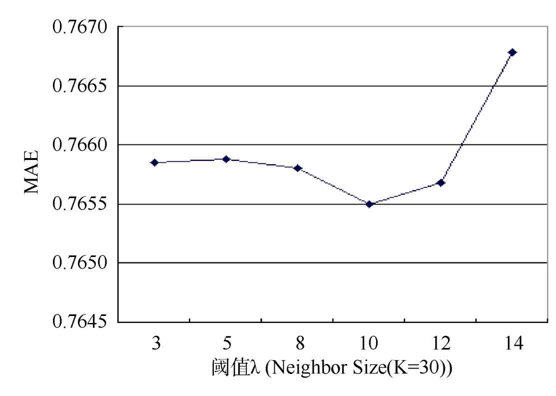


图 1　阈值 λ 对 MAE 值的影响

从图 1 中能够看出阈值 λ 取 10 时，MAE 值最低，推荐精度最高。因此在实验中设置阈值 λ =10。

(2) 实验二：参数 α 对推荐系统性能的影响

该实验主要用于测试参数 α 对 MAE 值的影响，同样需要控制最近邻的个数。当最近邻个数为 30 时，MAE 值随不同的 α 值的变化情况如图 2 所示：

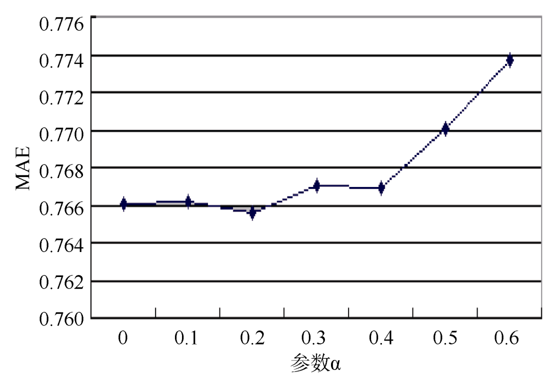


图 2　变量 α 对 MAE 的影响

从图 2 中可以看出，当 $\alpha = 0.2$ 时，MAE 值最小，推荐结果最优。随着 α 逐步增加至 0.2 时，MAE 值逐步减小，α 继续递增，MAE 值又开始缓慢递增。通过实验结果可以认为 α 在协同过滤中起着重要的作用，

①http://grouplens.org/datasets/movielens.

只有选择恰当的α才能获得最佳的推荐对象，得到最优的推荐结果，MAE值才能降到最低。所以在实验三中参数α的取值为0.2。

(3) 实验三：推荐性能随最近邻数目的变化情况

为了验证本文提出的基于用户评分时间改进的协同过滤推荐算法的有效性，进行实验对比。在MovieLens数据集上对改进的算法与传统的基于用户的协同过滤推荐算法的推荐准确性进行了比较。从实验一和实验二的结果可知，λ的取值为10，α的取值为0.2。MAE值随不同的最近邻数目的变化情况如图3所示：

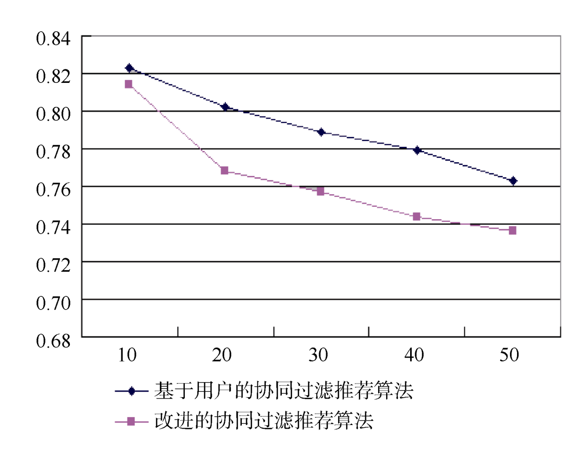


图3　改进后协同过滤算法的性能

从图3中可以看出，基于用户评分时间改进的协同过滤推荐算法在最近邻用户个数分别为10、20、30、40、50的MAE值均小于传统的基于用户的协同过滤推荐算法。MAE值平均降低2%。由此可见，在相似度的计算中考虑用户的评分时间、并且引入用户评分方差相似性后推荐效果的准确度得到了明显提高。

5　结　语

本文对传统的基于用户的协同过滤推荐算法中存在的不足之处进行改进，提出一种基于用户评分时间改进的协同过滤推荐算法。新算法考虑了在用户–项目评分矩阵异常稀疏、系统中两个用户之间共同评分项目极少时，所导致的相似度计算不准确，推荐准确性下降的情况。针对这一问题，本文结合用户的评分时间发现具有相似评分行为的用户，并融合用户评分方差相似度来改善传统用户之间相似度的计算，进而优化了目标用户最近邻的查找方式。实验结果表明，该算法即使在系统数据异常稀疏、用户之间共同评分稀少的前提下依然能够相对准确地计算用户之间的相似度，从而得到更加准确的推荐效果。

参考文献：

[1] 张莉，秦桃，腾丕强. 一种改进的基于用户聚类的协同过滤算法[J]. 情报科学, 2014, 32(10): 24-27. (Zhang Li, Qin Tao, Teng Piqiang. An Improved Collaborative Filtering Recmmendatiion Algorithm Based on User Clustering [J]. Information Science, 2014, 32(10): 24-27.)

[2] 方耀宁，郭云飞，丁雪涛，等. 一种基于局部结构的改进奇异值分解推荐算法[J]. 电子与信息学报，2013，35(6): 1284-1289. (Fang Yaoning, Guo Yunfei, Ding Xuetao, et al. An Improved Singular Value Decomposition Recommender Algorithm Based on Local Structures [J]. Journal of Electronics & Information Technology, 2013, 35(6): 1284-1289.)

[3] 孙辉，马跃，杨海波，等. 一种相似度改进的用户聚类协同过滤推荐算法[J]. 小型微型计算机系统，2014，35(9): 1967-1970. (Sun Hui, Ma Yue, Yang Haibo, et al. Collaborative Filtering Recommendation Algorithm by Optimizing Similarity and Clustering Users [J]. Journal of Chinese Computer Systems, 2014, 35(9): 1967-1970.)

[4] 高翔. 电子商务个性化推荐系统中协同过滤算法的研究[D]. 南京：南京航空航天大学, 2011. (Gao Xiang. Research of Collaborative Filtering on Recommendation Systems for E-Commerce [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2011.)

[5] 许智宏，王宝莹. 基于项目综合相似度的协同过滤算法[J]. 计算机应用研究, 2014, 31(2): 398-400. (Xu Zhihong, Wang Baoying. Collaborative Filtering Recommendation Algorithm Based on Item Complex Similarity [J]. Application Research of Computers, 2014, 31(2): 398-400.)

[6] 文俊浩，舒珊. 一种改进相似性度量的协同过滤推荐算法[J].计算机科学, 2014, 41(5): 68-71. (Wen Junhao, Shu Shan. Improves Collaborative Filtering Recommendation Algorithm of Similarity Measure [J]. Computer Science, 2014, 41(5): 68-71.)

[7] 严冬梅，鲁城华. 基于用户兴趣度和特征的优化协同过滤推荐[J]. 计算机应用研究，2012，29(2): 497-500. (Yan Dongmei, Lu Chenghua. Optimized Collaborative Filtering Recommendation Algorithm Based on Users' Interest Degree and Feature [J]. Application Research of Computers, 2012, 29(2): 497-500.)

[8] 赵雪. 基于用户兴趣的个性化协同过滤推荐算法研究[D]. 秦皇岛: 燕山大学, 2014. (Zhao Xue. The Personalized Collaborative Filtering Recommendation Algorithm Based on User Interest [D]. Qinhuangdao: Yanshan University, 2014.)

作者贡献声明:

李道国: 提出研究思路, 设计研究方案;
李连杰, 申恩平: 分析数据, 进行试验;
李道国, 李连杰: 论文起草及最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。
[1] 李道国, 李连杰, 申恩平. base.base. 基础数据集.
[2] 李道国, 李连杰, 申恩平. test.test. 测试数据集.

收稿日期: 2016-04-22
收修改稿日期: 2016-05-24

New Collaborative Filtering Recommendation Algorithm Based on User Rating Time

Li Daoguo[1] Li Lianjie[2] Shen Enping[2]
[1](School of Information Engineering, Hangzhou Dianzi University, Hangzhou 310018, China)
[2](School of Management, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: [**Objective**] This paper tries to solve the problems facing traditional collaborative filtering algorithm due to sparse data and few users' common scores, and then improve the accuracy of the score prediction systems. [**Methods**] First, we identified users with similar scoring behaviors based on their scoring time. Second, we integrated the similarity of user score variance to the calculation of similarity. [**Results**] The new algorithm, which reduced the MAE by 2% compared to the traditional algorithm, improved the performance of recommendation system. [**Limitations**] The proposed algorithm was only examined with the MovieLens dataset, which needed to be expanded to other datasets. [**Conclusions**] The proposed algorithm can improve the effectiveness of recommendation systems.

Keywords: Collaborative filtering Data sparsity Similarity score User rating variance similarity Nearest neighbor